

Spatially Correlated Content Caching for Device-to-Device Communications

Derya Malak, Mazin Al-Shalash, and Jeffrey G. Andrews

Abstract

We study optimal geographic content placement for device-to-device (D2D) networks in which each file's popularity follows the Zipf distribution. The locations of the D2D users (caches) are modeled by a Poisson point process (PPP) and have limited communication range and finite storage. We are interested in the optimal placement of content into the caches, which intuitively should not be spatially independent, since if the file is already cached nearby, it is less useful to cache the file again. We propose several novel spatially correlated caching strategies, and contrast them with the baseline independent content placement which is used in most prior work. These include (i) a spatially exchangeable content placement technique to prioritize the caches for content placement and (ii) the “Gibbs” point process-based soft-core caching model that captures the pairwise interactions between users, including attraction or repulsion. We derive and optimize the *hit probability*, which is the probability that a given D2D node can find a desired file at another node within its communication range. Exchangeable placement actually performs worse than independent placement, while the Gibbs model shows that repulsive cache placement often yields a higher hit probability. We consider two further special cases of Gibbs repulsive cache placement: (ii.a) the Fermi-Dirac and (ii.b) Matérn hard-core (MHC) models, and show that MHC placement is effective for small cache sizes and a small communication radius, which are likely conditions for D2D.

I. INTRODUCTION

D2D communication is a promising technique for enabling proximity-based applications and increased offloading from the heavily loaded cellular network, and is being actively standardized

This work in part appeared in Proc. IEEE Intl. Symposium on Info. Theory, Barcelona, Spain, July 2016 [1].

D. Malak and J. G. Andrews are with the Wireless Networking and Communications Group (WNCG), The University of Texas at Austin, Austin, TX 78701 USA (email: deryamalak@utexas.edu; jandrews@ece.utexas.edu).

M. Al-Shalash is with Huawei Technologies, Plano, TX 75075 USA (e-mail: mshalash@huawei.com). Last revised: September 5, 2016.

by 3GPP [2]. Its efficacy requires users to possess content that a nearby user wants. Therefore, intelligent caching of popular files is necessary for D2D to be successful. Caching has been shown to provide increased spectral reuse and throughput gain in D2D-enabled networks [3], but the optimal way to spatially cache content is unknown. Intuitively, given a finite amount of storage at each node, popular content should be seeded into the network in a way that maximizes the *hit probability* that a given D2D device can find a desired file – selected at random according to a request distribution – within its radio range. We explore this problem quantitatively in this paper by considering different spatial content models and deriving, optimizing and comparing the hit probabilities for each of them.

A. Related Work

Content caching has received significant recent attention as a means of improving the throughput and latency of networks without requiring additional bandwidth or other technological improvements. A practical use case of caching is video, which will consume nearly 70% of all wireless data by 2019 [4]. Video caching appears particularly profitable and plausible compared to other types of content [5], and is perfectly suited to D2D networks for offloading traffic from congested cellular networks.

Research to date on content caching has been mainly focused on two different perspectives. On one hand, researchers have attempted to understand the fundamental limits of caching. The gain offered by local caching was characterized in the landmark paper by Maddah-Ali and Niesen [6], while scaling laws for the number of active D2D links and optimal collaboration distance with D2D content caching are studied in [7], [8]. Alternatively, several decentralized caching studies have optimized the caching distribution to determine the optimal association of content to wireless caches in order to maximize the cache hit probability, given a base station (BS)-user topology, or in stochastic settings such as [9], [10], [11]. There are also geographic placement models such as [12], [13], [14] that exploit stochastic geometry to maximize the cache hit probability where various point processes are used to model the cache locations. However, most of these strategies suggest caching the most popular content only and do not exploit the diversity of content, that is, they are agnostic to the files available in nearby locations. As the current paper will show, however, it is not usually optimal to cache just the most popular files. Furthermore, for larger transmission range and higher network density, we will quantify and

see that the hit-maximizing caching strategy is increasingly skewed away from only caching the most popular files.

A related line of work is the proactive caching – also known as “pre-fetching” – of high bandwidth content at the network edge [15], [5]. Pre-fetching can smooth out the network traffic over time to avoid peaks in space and time, which reduces data delivery costs. Proactive caching can occur at the end user [16] or at a small cell BSs to reduce peak demands on the backhaul link, which is often the bottleneck [11], [17], [18]. Despite their usefulness, none of these approaches capture the spatial interactions in the network or provide guidance on how the cache states at nearby nodes should effect the caching strategy at a given node.

Challenges for the adoption of caching for wireless access networks also include making timely estimates of varying content popularity [19]. Cache update algorithms exploiting the temporal locality of the content have been well studied [20]. Inspired from the Least Recently Used (LRU) replacement principle, a multi-coverage caching policy at the edge-nodes is proposed in [21], where caches are updated in a way that provides content diversity to users who are covered by more than one node. Although [21] combines the temporal and spatial aspects of caching and approaches the performance of centralized policies, it is restricted to the LRU principle.

To the best of our knowledge, a decentralized content caching policy that captures the joint spatial interactions of the nodes has not been studied in the literature. We aim to maximize the cache hit probability for a D2D communication network where the spatial distribution of nodes can be exploited to efficiently use the caches. The goal is to decide where to store files as a function of the user locations.

B. Contributions

We consider a D2D network in which the D2D user locations are modeled by a Poisson point process (PPP), and users have limited communication range and finite storage. The D2D users are served by each other if the desired content is cached at a user within its radio range: this is called a *hit*. Otherwise, they are served by the cellular network base station, which is what D2D communication aims to avoid. We wish to optimize the cache hit probability.

Spatial caching. We introduce a spatial content distribution model for a D2D network in Sect. II, and describe the cache hit probability maximization problem in Sect. III. We detail independent content placement (no spatial correlation) in Sect. III-B. We then propose novel

spatially correlated cache placement strategies that enable spatial diversity to maximize the D2D cache hit probability: (i) a spatially exchangeable content placement technique to prioritize the caches for content placement, is introduced in Sect. IV, (ii) general Gibbs point process-based caching models as detailed in Sect. V, including the special cases of (ii.a) the Fermi-Dirac in Sect. VI and (ii.b) Matérn hard-core models in Sect. VII. We show that exchangeable placement yields a positively correlated spatial distribution of content, and is suboptimal in terms of the cache hit probability compared to independent placement. However, as the coverage number [22] –the number of transmitters simultaneously covering a user– increases, the performance of exchangeable placement approaches the performance of independent placement.

Pairwise interactions and Gibbs distribution-based placement. Inspired from Ising models characterizing the pairwise interactions of a node with its neighbors, we devise a content placement strategy using Gibbs point processes (GPPs) that are mathematical models (variants of Ising models) of particle interactions in statistical mechanics. GPPs are characterized by a potential function, modeling the interactions –e.g., attraction or repulsion– among nodes. We consider a general GPP-based cache placement in Sect. V, and show that repulsive cache placement yields a higher hit probability. Then, we consider some specific examples of repulsive cache placement, such as permutation distribution and Fermi-Dirac (F-D) distribution, in Sect. VI.

Matérn hard-core-based placement. The GPP-based models proposed in Sects. V and VI with repulsive potentials generate soft-core placement models. In the case of a soft-core point process, thinning is stronger the closer point pairs of the initial PPP are, but any pair distance still has non-vanishing probability. In the case of hard-core processes, after thinning, points can not exist within a certain radius of each point in the thinned point process. In the Matérn hard-core (MHC) model, the caches storing a particular file are never closer to each other than some given distance, meaning that MHC placement yields a negatively correlated spatial content distribution, so neighboring users are not likely to cache redundant content. In Sect. VII, we consider two different models for MHC-based placement: (i) MHC-A that provides a higher cache hit probability than the independent placement in the small cache size regime and (ii) MHC-B that provides a higher cache hit probability than the independent placement for short range communication.

Comparisons and design insights. Sect. VIII provides a detailed simulation study to compare the performance between the different content placement strategies. Our results show that geographic placement should capture locality of content. Independent content placement does not exploit D2D interactions at the network level. Exchangeable placement, being correlated, performs worse than independent content placement. Negatively correlated placement, e.g., permutation distribution, or more generally statistical physics models, like Gibbs and Ising models can exploit locality. Models perform differently in different communication and cache size regimes. For small communication ranges and small cache sizes, MHC placement is preferred, and for large communication ranges and large cache sizes, independent placement model is a better alternative.

II. SYSTEM MODEL

The locations of the D2D users are modeled by a PPP Φ with density λ as in [23]. We assume that there are M total files in the network and each user has a cache of finite size $N < M$. Depending on its cache state, each user makes requests for new files based on a general popularity distribution over the set of the files. The popularity of such requests is modeled by the Zipf distribution, which has pmf $p_r(n) = \frac{1}{n^{\gamma_r}} / \sum_{m=1}^M \frac{1}{m^{\gamma_r}}$, for $n = 1, \dots, M$, where γ_r is the Zipf exponent that determines the skewness of the distribution.

D2D users can only communicate within a finite range, which we call the D2D radius, denoted by R_{D2D} . A file request is fulfilled by the D2D users within the D2D radius if one has the file, else, the D2D user is served by the cellular network. The coverage process of the proposed model can be represented by a Boolean model as described next.

Definition 1. *The Boolean model (BM) is based on a PPP, whose points are also called germs, and on an independent sequence of iid compact sets called grains, defined as a model driven by an independently marked PPP on \mathbb{R}^2 [24].*

Consider a given realization $\phi = \{x_i\} \subset \mathbb{R}^2$ of the PPP Φ . We can think of ϕ as a counting measure or a point measure $\phi = \sum_i \delta_{x_i}$, $x_i \in \mathbb{R}^2$, where x_i denotes the coordinates of the i^{th} user and $\delta_x = \{0, 1\}$ is the Dirac measure at x ; for $A \subset \mathbb{R}^2$, $\delta_x(A) = 1$ if $x \in A$ and $\delta_x(A) = 0$ if $x \notin A$. Consequently, $\phi(A)$ gives the number of points of ϕ in A .

Our model is a simple BM where x_i 's denote the germs and $B_i(R_{D2D})$ is a closed ball of radius R_{D2D} centered at x_i , which denotes the grains. Then, the coverage process is driven by

the following independently marked PPP: $\tilde{\Phi} = \sum_i \delta_{(x_i, B_i(R_{D2D}))}$. The BM is given by the union $V_{BM} = \bigcup_i (x_i + B_0(R_{D2D}))$ that models the coverage process of the D2D transmitters.

Definition 2. Volume fraction [24]. *Since our model is translation invariant, the volume fraction can be expressed as the probability that the origin is covered by $B_0(R_{D2D})$ given by*

$$p = \mathbb{P}(0 \in B_0(R_{D2D})) = 1 - \exp(-\lambda \pi R_{D2D}^2). \quad (1)$$

We propose different strategies to serve the D2D requests that maximize the cache hit probability. Assuming a transmitter receives one request at a time and multiple transmitters can potentially serve a request, the selection of an active transmitter depends on the caching strategy. A summary of the symbol definitions and important network parameters are given in Table I.

III. AVERAGE CACHE HIT PROBABILITY

To characterize the successful transmission probability, one needs to know the number of users that a typical node can connect to, i.e., the coverage number. Exploiting the properties of the PPP, the distribution of the number of transmitters covering the typical receiver that requests file m is given by

$$\mathcal{N}_m \sim \text{Poisson}(\lambda_{t,m} \pi R_{D2D}^2). \quad (2)$$

Assume that the files are cached at the D2D users identically and independently of each other and let $p_c(\cdot)$ be the caching probability. Let Y_m be the indicator random variable that takes the value 1 if file m is available in the cache and 0 otherwise. Thus, the caching probability of file m is given by $p_c(m) = \mathbb{P}(Y_m = 1)$. Any cache satisfies the condition $\sum_{m=1}^M Y_m \leq N$, i.e., Y_m 's are inherently dependent. However, for tractability reasons¹ and due to the independent content placement assumption, we take the expectation of this relation and obtain our cache constraint: $\sum_{m=1}^M \mathbb{P}(Y_m = 1) = \sum_{m=1}^M p_c(m) \leq N$.

The maximum average total cache hit probability, i.e., the probability that the typical user

¹The original problem is combinatorial and is NP-hard.

Symbol	Definition
D2D communication radius	R_{D2D}
Density of D2D users; Density of D2D users having file m	$\lambda_t; \lambda_{t,m}$
Density of requesting users; Density of users requesting file m	$\lambda_r; \lambda_{r,m}$
Number of D2D users covering a user requesting a file	$\mathcal{N} \sim \text{Poisson}(\lambda_t)$
Number of D2D users covering a user requesting file m	$\mathcal{N}_m \sim \text{Poisson}(\lambda_{t,m})$
Hit probability for placement strategy X	$P_{\text{Hit},X}$
Miss probability of file m given k users cover the typical receiver for placement strategy X	$P_{\text{Miss},X}(m, k)$
Zipf exponent for request distribution	γ_r
File request distribution; file caching distribution	$p_r(\cdot) \sim \text{Zipf}(\gamma_r); p_c(\cdot)$
Total number of files; cache size	$M; N < M$
Cache design parameters for independent content placement	$L < N; K > N$
The set of files that should be stored with probability 1	$\{1, \dots, L-1\}$
The set of files that should be discarded with probability 1	$\{K+1, \dots, M\}$
Exclusion radius of file m for the MHC model	r_m
Ball centered at x with radius R_{D2D}	$B_x(R_{\text{D2D}})$
k dimensional bounded region $[0, D]^k$	D^k

TABLE I: Notation.

finds the content in one of the D2D users it is covered by, can be evaluated by solving

$$\begin{aligned}
 & \max_{p_c} P_{\text{Hit},I} \\
 & \text{s.t.} \quad \sum_{m=1}^M p_c(m) \leq N,
 \end{aligned} \tag{3}$$

where $P_{\text{Hit},I} = 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}_m = k)(1 - p_c(m))^k$. First, we consider the following trivial case.

Proposition 1. Caching most popular content only. *The baseline solution is to store the most popular contents only. Letting $Y_m = 1_{m \leq N}$, the miss probability is $P_{\text{Miss},I}(m, k) = 1_{N < m \leq M}$ for any coverage number k . The average hit probability is $P_{\text{Hit},I} = 1 - \sum_{m=N+1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}_m = k) = \sum_{m=1}^N p_r(m)$.*

A. Independent Cache Design

Optimal content placement is a binary problem satisfying $\sum_{m=1}^M Y_m = N$. However, as noted above, the constraint in (3) is based on the average values of Y_m 's, which yields a relaxed content

placement. Later, we show there are feasible solutions to the relaxed problem filling up all the cache slots.

The key step in evaluating (3) is to determine the coverage number distribution, i.e., $\mathbb{P}(\mathcal{N}_m = k)$. We can optimize $P_{\text{Hit},l}$ by using the Lagrangian technique as follows

$$\mathcal{L}(\mu) = 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}_m = k) (1 - p_c(m))^k - \mu \left(\sum_{m=1}^M p_c(m) - N \right).$$

Taking the derivative of $\mathcal{L}(\mu)$ with respect to $p_c(m)$ and evaluating at $\mu = \mu^*$, we have $\left. \frac{d\mathcal{L}(\mu)}{dp_c(m)} \right|_{\mu=\mu^*} = 0$, for which there exists a feasible solution $p_c^*(m)$ that satisfies

$$p_r(m) \sum_{k=1}^{\infty} k \mathbb{P}(\mathcal{N}_m = k) (1 - p_c^*(m))^{k-1} = \mu^*, \quad (4)$$

where $p_r(m) \mathbb{P}(\mathcal{N}_m = 1) \leq \mu^* \leq p_r(m) \mathbb{E}[\mathcal{N}_m]$.

Similar to the approach in [12], we can use the bisection method² and numerically solve (4) to find the $p_c^*(m)$ values. We initialize the bisection method by setting μ such that $\mu \in [\mu_{\min}, \mu_{\max}]$, where $\mu_{\max} = p_r(N/c_b) \mathbb{P}(\mathcal{N}_{N/c_b} = 1)$ assuming $p_c(m) = 1$ for $m \leq N/c_b$, hence $\mu^* \leq \mu_{\max}$, and $\mu_{\min} = p_r(c_b N) \mathbb{E}[\mathcal{N}_{c_b N}]$ assuming $p_c(m) = 0$ for $m \geq c_b N$, hence $\mu^* \geq \mu_{\min}$. Here, c_b is a constant integer parameter appropriately adjusted and N is divisible by c_b and $c_b N \leq M$.

Using the coverage number pmf (2), we can rewrite (4) as

$$\mu^* = p_r(m) \sum_{k=1}^{\infty} k e^{-\lambda_{t,m} \pi R_{\text{D2D}}^2} \frac{(\lambda_{t,m} \pi R_{\text{D2D}}^2)^k}{k!} (1 - p_c^*(m))^{k-1} = p_r(m) \lambda_{t,m} \pi R_{\text{D2D}}^2 e^{-p_c^*(m) \lambda_{t,m} \pi R_{\text{D2D}}^2},$$

which yields the following expression

$$p_c^*(m) = \begin{cases} 1 & \mu^* \leq p_r(m) \mathbb{P}(\mathcal{N}_m = 1) \\ \frac{1}{\lambda_{t,m} \pi R_{\text{D2D}}^2} \log \left(\frac{p_r(m) \lambda_{t,m} \pi R_{\text{D2D}}^2}{\mu^*} \right) & \mu^* \in \mathcal{M}_m \\ 0 & \mu^* \geq p_r(m) \mathbb{E}[\mathcal{N}_m] \end{cases}, \quad (5)$$

where $\mathbb{P}(\mathcal{N}_m = 1) = e^{-\lambda_{t,m} \pi R_{\text{D2D}}^2} (\lambda_{t,m} \pi R_{\text{D2D}}^2)$, $\mathbb{E}[\mathcal{N}_m] = \lambda_{t,m} \pi R_{\text{D2D}}^2$ and \mathcal{M}_m is a set such that for any $\mu^* \in \mathcal{M}_m$, it is satisfied that $p_r(m) \mathbb{P}(\mathcal{N}_m = 1) \leq \mu^* \leq p_r(m) \mathbb{E}[\mathcal{N}_m]$.

²The bisection method is a numerical root-finding method that repeatedly bisects an interval and selects a subinterval in which a root must lie. The algorithm stops when the change in the root is smaller than a chosen $\varepsilon > 0$.

B. A Linear Approximation to Independent Cache Design

Given that each cache can store $N < M$ files³, our objective is to determine the number of files L that should be stored in the cache with probability 1, and the maximum number of distinct files K that can be stored in the cache as a function of the important design parameters, e.g., R_{D2D} , $\{\lambda_{t,m}\}$ and N . Incorporating the finite cache size constraint to (5), we can rewrite $\sum_{m=1}^M p_c(m)$ as follows:

$$L - 1 + \sum_{m=L}^K \frac{1}{\lambda_{t,m} \pi R_{D2D}^2} \log \left(\frac{p_r(m) \lambda_{t,m} \pi R_{D2D}^2}{\mu^*} \right) = N. \quad (6)$$

Using the boundary conditions for μ^* , we have $p_r(K) \lambda_{t,K} \pi R_{D2D}^2 \leq \mu^* \leq p_r(L) e^{-\lambda_{t,L} \pi R_{D2D}^2} (\lambda_{t,L} \pi R_{D2D}^2)$, where the relation between L and K can be found as $p_r(K)^2 \leq p_r(L)^2 \exp(-\lambda_{t,L} \pi R_{D2D}^2)$, which follows from $\lambda_{t,m} = \lambda p_r(m)$, i.e., the density of the transmitting users is proportional to the density of the requests. Using (5), for any $L \leq m \leq K$, the optimal solution is

$$p_c^*(m) = \frac{2\gamma_r}{\lambda \pi R_{D2D}^2} \log \left(\frac{K}{m} \right) m^{\gamma_r} \sum_{j=1}^M \frac{1}{j^{\gamma_r}} + \left(\frac{m}{K} \right)^{\gamma_r} p_c(K). \quad (7)$$

From (6) and (7), we obtain the following relation:

$$N - L + 1 = \left[\frac{\sum_{j=1}^M (2\gamma_r / j^{\gamma_r})}{\lambda \pi R_{D2D}^2} \log(K) + \frac{p_c(K)}{K^{\gamma_r}} \right] \sum_{m=L}^K m^{\gamma_r} - \frac{\sum_{j=1}^M (2\gamma_r / j^{\gamma_r})}{\lambda \pi R_{D2D}^2} \sum_{m=L}^K \log(m) m^{\gamma_r}. \quad (8)$$

Applying $p_r(K)^2 \leq p_r(L)^2 \exp(-\lambda_{t,L} \pi R_{D2D}^2)$ with equality and from (8), we uniquely determine L and K that approximate the optimal content placement pmf in (5) as the following linear model:

$$p_c^{\text{Lin}}(m) = \min \left\{ 1, \left(1 - \frac{m - L}{K - L} \right)^+ \right\}, \quad (9)$$

where $y^+ = \max\{y, 0\}$, which is a good approximation as shown in Sect. VIII.

³Swapping the contents within a cache does not change cache's state.

IV. SPATIALLY EXCHANGEABLE CACHE PLACEMENT

From a user's perspective, the exact location of cached content is not important as long as it is available within R_{D2D} . This is illustrated in Fig. 1 by an example with two equivalent models. In both models, the number of caches having any content type is the same. However, the locations where the content is cached are different. More generally, from the typical user's perspective, any finite permutation of any content type among the caches within R_{D2D} of the user is equivalent.

In this section, we assume a spatially exchangeable cache model defined as follows. For an ordered set of n transmitters covering a typical receiver with desired content m , the binary sequence $Y_{m_1}, Y_{m_2}, \dots, Y_{m_n}$ denotes the availability of the content in the respective caches:

$$Y_{m_i} = \begin{cases} 1, & \text{file } m \text{ is available in cache } i, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

We assume the sequence $Y_{m_1}, Y_{m_2}, \dots, Y_{m_n}$ is *exchangeable* in the spatial domain.

Definition 3. Exchangeable random variables. *An exchangeable sequence Y_1, Y_2, \dots of random variables is such that for any finite permutation r of the indices $1, 2, \dots$, the joint probability distribution of the permuted sequence $Y_{r(1)}, Y_{r(2)}, \dots$ is the same as the joint distribution of the original sequence.*

A theoretical description of exchangeability is given now.

Theorem 1. de Finetti's theorem. *A binary sequence Y_1, \dots, Y_n, \dots is exchangeable if and only if there exists a distribution function F on $[0, 1]$ such that for all n $p(y_1, \dots, y_n) = \int_0^1 \theta^{t_n} (1 - \theta)^{n-t_n} dF(\theta)$, where $p(y_1, \dots, y_n) = \mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n)$ is the joint pmf and $t_n = \sum_{i=1}^n y_i$. It further holds that F is the distribution function of the limiting frequency, i.e., if $X = \lim_{n \rightarrow \infty} \sum_i Y_i / n$ a.s., then $\mathbb{P}(X \leq x) = F(x)$ and by conditioning with $X = \theta$, we obtain*

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n | X = \theta) = \theta^{t_n} (1 - \theta)^{n-t_n}. \quad (11)$$

The formulation to maximize the cache hit for an exchangeable content placement strategy

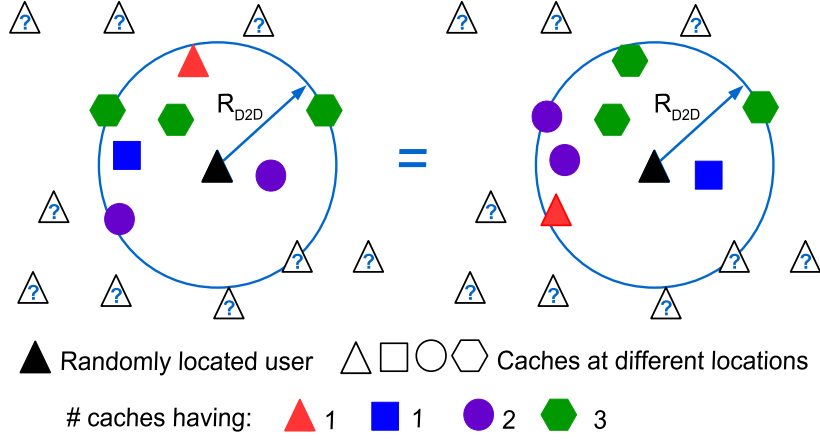


Fig. 1: Exchangeable cache placement with two equivalent models, where the same set and multiplicity of files are permuted among the caches within R_{D2D} of the randomly located user.

becomes

$$\begin{aligned}
 & \max_{f_{X_m}} P_{\text{Hit},E} \\
 & \text{s.t.} \quad \int_0^1 dF_{X_m}(\theta) = 1, \quad m \in \{1, \dots, M\} \\
 & \quad \sum_{m=1}^M \mathbb{E}[X_m] \leq N,
 \end{aligned} \tag{12}$$

where $P_{\text{Hit},E} = 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}_m = k) P_{\text{Miss},E}(m, k)$, and $P_{\text{Miss},E}(m, k)$ is the probability that k caches cover a receiver, and none has file m . The constraints are such that the distribution functions F_{X_m} for $m \in \{1, \dots, M\}$ are on $[0, 1]$, and $\mathbb{E}[X_m] = \int_0^1 \theta f_{X_m}(\theta) d\theta$ is the probability a cache contains file m and each cache contains N files in total on average⁴.

From Theorem 1, the average cache miss probability $P_{\text{Miss},E}(m, k)$ is given by

$$P_{\text{Miss},E}(m, k) = \int_0^1 (1 - \theta)^k f_{X_m}(\theta) d\theta = \mathbb{E}[(1 - X_m)^k]. \tag{13}$$

Hence, the objective in (12) is equal to

⁴Same as the relaxed constraint in the formulation (3).

$$\begin{aligned}
P_{\text{Hit,E}} &= \sum_{m=1}^M p_r(m) \int_0^1 \left[1 - \sum_{k=0}^{\infty} \exp(-\lambda_{t,m} \pi R_{\text{D2D}}^2) \frac{(\lambda_{t,m} \pi R_{\text{D2D}}^2)^k}{k!} (1-\theta)^k \right] f_{X_m}(\theta) d\theta \\
&= 1 - \sum_{m=1}^M p_r(m) \mathbb{E}[\exp(-\lambda_{t,m} \pi R_{\text{D2D}}^2 X_m)].
\end{aligned} \tag{14}$$

Future samples behave like earlier samples, meaning formally that any order (of a finite number of samples) is equally likely. This formalizes the notion of the future being predictable on the basis of past experience. To give more intuition on exchangeability, we next consider an example.

Example 1. Sampling Without Replacement. *A simple strategy to generate a set of random variables that satisfies exchangeability is as follows. Fix the number transmitters n covering a receiver with desired content m , and consider any permutation $Y_{r(m_1)}, \dots, Y_{r(m_n)}$. Conditionally place the content to cache: $\mathbb{P}(Y_{r(m_k)} = 0 | Y_{r(m_1)} = 0, Y_{r(m_2)} = 0, \dots, Y_{r(m_{k-1})} = 0) = \frac{k}{k+1}$ for $1 \leq k \leq n$. Hence, the miss probability for file m given k caches cover a receiver is $P_{\text{Miss,E}}(m, k) = \frac{1}{2} \times \frac{2}{3} \times \dots \times \frac{k}{k+1} = \frac{1}{k+1}$.*

This example is a special case of the de Finetti's Theorem where the limiting random variables are uniformly distributed on $[0, 1]$, i.e., $X_m \sim F_m = U[0, 1]$ for $m \in \{1, \dots, M\}$. Hence, from (13),

$$P_{\text{Miss,E}}(m, k) = \int_0^1 (1-\theta)^k d\theta = \frac{1}{k+1}.$$

Proposition 2. *Any exchangeable content placement strategy is worse than independent placement in terms of the average cache hit probability.*

Proof: Using the convexity of exponential function, we rewrite the objective of (14) as follows:

$$1 - P_{\text{Hit,E}} = \sum_{m=1}^M p_r(m) \mathbb{E}[\exp(-\lambda_{t,m} \pi R_{\text{D2D}}^2 X_m)] \stackrel{(a)}{\geq} \sum_{m=1}^M p_r(m) \exp(-\lambda_{t,m} \pi R_{\text{D2D}}^2 \mathbb{E}[X_m]), \tag{15}$$

Hence, the miss probability of the exchangeable placement model is higher than the miss probability of the independent placement (a), where $p_c(m) = \mathbb{E}[X_m]$ is the independent placement probability. ■

The next result is a more generalized version of Proposition 2 in the sense that it holds for any kind of coverage distribution $\mathbb{P}(\mathcal{N}_m = \cdot)$.

Lemma 1. *Given any coverage distribution, which include the Boolean model and the Signal-to-Interference-and-Noise-Ratio (SINR) model or any other coverage model, the exchangeable placement strategy always performs worse than the independent placement strategy.*

Proof: The hit probability for the exchangeable model is given by

$$P_{\text{Hit,E}} \stackrel{(a)}{=} 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}_m = k) \mathbb{E}[(1 - X_m)^k] \stackrel{(b)}{\leq} 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}_m = k) (1 - \mathbb{E}[X_m])^k,$$

where X_m 's are the limiting random variables, (a) follows from (13), i.e., $P_{\text{Miss,E}}(m, k) = \mathbb{P}(\bigcap_{l=1}^k \{Y_m = 0\}) = \mathbb{E}[(1 - X_m)^k]$, and from exchangeability, the distribution function of X_m , i.e., F_{X_m} is on $[0, 1]$, hence (b) follows from the convexity of $(1 - X_m)^k$ for $k \in \mathbb{Z}_{\geq 0}$, i.e., $P_{\text{Miss,E}}(m, k) \geq (1 - \mathbb{E}[X_m])^k$. Denoting the caching probability of file m for the independent placement model by $\mathbb{E}[X_m] = \mathbb{P}(Y_m = 1)$, from (13), it is clear that $P_{\text{Miss,E}}(m, k) \geq \mathbb{P}(Y_m = 0)^k$, and the independent placement strategy gives a higher average hit probability compared to the exchangeable strategy. ■

A wide class of random processes exhibit exchangeability, which include combinatorial stochastic processes, Markov chains, coalescent processes, Poisson-Dirichlet processes, Erdős-Rényi graphs, the Indian buffet process, the Chinese restaurant process, and a large collection of statistical mechanical systems on complete graphs. Interested reader can refer to [25] for further examples.

V. GIBBS DISTRIBUTION-BASED CACHE PLACEMENT

Gibbs point processes (GPPs) are mathematical models of particle interactions in statistical mechanics. GPPs form one of the most important classes of point processes in spatial statistics that may incorporate dependence between the points [26, Ch. 6]. They are used to model the pairwise interaction processes, one of the building blocks of modern statistical physics [27]. GPPs are not a universal class of models which can be used in all situations. They are good models for patterns with some degree of regularity, i.e., more regular than Matérn hard-core (MHC) processes as we detail in this section, or for moderate clustering, but can be deficient in cases of strong clustering [28, Ch. 5.5].

GPPs are related to the so-called Boltzmann-Gibbs distributions, which describe equilibrium states of subsystems of very large closed physical systems. Special cases include the Ising model [29], Boltzmann distribution or the Fermi-Dirac model, Markov point processes, cluster processes such as the Neyman-Scott processes, and repulsive processes such as the Strauss model, and hard-core point processes. Spatial birth-and-death processes have a close relationship with GPPs [28, Ch. 5.5].

A. Background on Gibbs Point Processes

Consider a GPP Φ_G of distribution P with exactly k points in a bounded region (bounded set B), i.e., the number of points in the window of observation is fixed [28, Ch. 5.5]. Assume that the distribution of the point process is given by a probability density function $f : \mathbb{R}^{dk} \rightarrow [0, \infty)$ so that

$$P(\Phi_G \in Y) = \int \cdots \int_{\{x_1, \dots, x_k\} \in Y} f(x_1, \dots, x_k) dx_1 \dots dx_k, \quad \text{for } Y \in \mathcal{N}_{B,k}, \quad (16)$$

where $\mathcal{N}_{B,k}$ denotes the trace of \mathcal{N} on the set of all point processes with k points in B . Because point processes are an unordered set of points, $f(x_1, \dots, x_k)$ is taken not to depend on the order of the arguments. It is given by the following formula $f(x_1, \dots, x_k) = \exp(-E(x_1, \dots, x_k))/Z$, where Z is a normalization constant, which is called the configurational partition function, and the function $E : \mathbb{R}^{dk} \rightarrow \mathbb{R} \cup \{\infty\}$ is called the energy function or the multiparticle functional, which does not depend on the order of the arguments. These terms come from statistical mechanics [28, Ch. 5.5].

Pair potential function. The energy function E is frequently chosen as the following sum

$$E(x_1, \dots, x_k) = \beta \sum_{1 \leq i < j \leq k} \theta(\|x_i - x_j\|), \quad (17)$$

where $\theta : [0, \infty) \rightarrow (-\infty, \infty]$ is the pair potential, which also comes from statistical physics, and $\beta = T^{-1}$ is called the inverse temperature [28, Ch. 5.5]. Then, f takes the form

$$f(x_1, \dots, x_k) = \exp\left(-\beta \sum_{1 \leq i < j \leq k} \theta(\|x_i - x_j\|)\right)/Z. \quad (18)$$

The pair potential characterizes the GPP of density f constructed according to (18). A typical example is shown in Fig. 2. The potential $\theta(r)$ is infinite for $r \leq R$, i.e., the inter-node distance

can never be less than R . Therefore, the point process is in fact a hard-core model. For $r > R$, $\theta(r) = \exp\left(\frac{r}{r-R}\right) - 100 \exp\left(-\frac{r}{2} - R\right)$. Since $\theta(r)$ is large when r is slightly larger than R , such inter-point distances exist with a low probability. Inter-node distances for which $\theta(r)$ takes its minimum, i.e., the inter-point distances close to R_1 , should occur relatively frequently.

Examples of potential functions include the Mie potential [30] that models the short-range repulsive interactions and the longer range attractive interactions, Lennard-Jones potential [31], which is a special form of the Mie potential, and Morse/Long-range potential [32]. Parametric families of pair potentials θ are also studied [33], where $\theta(r) = b\left(\frac{\sigma}{r}\right)^q - a\left(\frac{\sigma}{r}\right)^s$, where $a \geq 0$, $b, \sigma > 0$ and $q > s$. This potential is a common model in statistical mechanics, called the Lennard-Jones potential. The case $n = 12$, $m = 6$ provides both types of interaction, attraction and repulsion, at different scales. For the purposes of the current paper, the potential function is assumed to be monotonic.

Partition function. Partition function describes the statistical properties of a system in thermodynamic equilibrium, and is a function of the state variables, such as the temperature and volume [28, Ch. 5.5]. It plays the role of a normalizing constant, ensuring that the state probabilities sum up to one. The partition function Z is difficult to compute [28, Ch. 5.5]. An approximation method is proposed in [33] for the case of pair potentials, and for a bounded region $[0, D]^2 \in \mathbb{R}^2$ with k particles, it is given as

$$Z_k \approx (\pi D^2)^k \left(1 - \frac{2\pi \int_0^D (1 - \exp(-\beta\theta(r)))rdr}{\pi D^2}\right)^{k(k-1)/2}. \quad (19)$$

B. Content Caching Exploiting Pairwise Interactions

A caching network modeled by a Gibbs distribution might not require centralized coordination since it captures the pairwise interactions among the nodes in a distributed manner. It is hard to characterize GPPs in their most generic form to optimize the performance of caching. In this section, we formulate the general hit probability maximization problem for the GPPs. Later in Sects. VI and VII, we consider special tractable cases of the model.

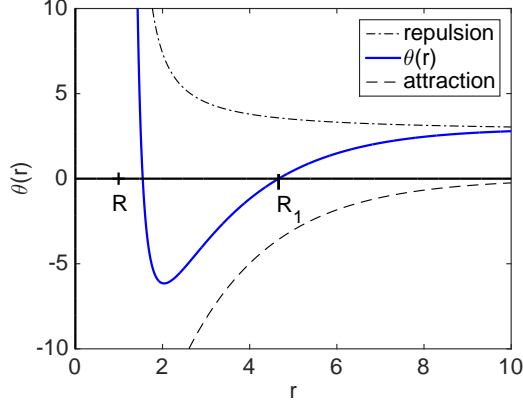


Fig. 2: A typical pair potential, the result of superposition of attractive and repulsive forces.

The maximum hit for the Gibbs content placement is given by the solution of:

$$\begin{aligned} \max_{\beta_m} \quad & P_{\text{Hit,G}} \\ \text{s.t.} \quad & \sum_{m=1}^M \mathbb{E}_{\mathcal{N}} [f_m(x_1, \dots, x_{N+1})] \leq N, \end{aligned} \quad (20)$$

where $P_{\text{Hit,G}} = 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N} = k) P_{\text{Miss,G}}(m, k)$, where $\mathbb{P}(\mathcal{N} = k) \sim \text{Poisson}(\lambda_t \pi D^2)$ and $P_{\text{Miss,G}}(m, k) = \int \cdots \int_{V^k} f_m(x_1, \dots, x_k) dx_1 \dots dx_k$. The constraint equation follows from that the probability that content m is cached at a transmitter is equal to $\mathbb{E} [f_m(x_1, \dots, x_{N+1})]$, and there are at most N files to be stored in each cache, and V^k characterizes the miss region given there exists k nodes, i.e., it is the k dimensional region $[0, D]^k \setminus [0, R_{\text{D2D}}]^k$.

$$\begin{aligned} f_m(x_1, \dots, x_k) &= \exp(-\beta_m E(x_1, \dots, x_k)) / Z_{m,k} \\ &= \exp\left(-\beta_m \sum_{1 \leq i < j \leq k} \theta(\|x_i - x_j\|)\right) / Z_{m,k}, \end{aligned} \quad (21)$$

where $Z_{m,k} = \int \cdots \int_{D^k} \exp(-\beta_m E(x_1, \dots, x_k)) dx_1 \dots dx_k$ is the partition function for index m given there are k nodes in the model, and D^k is the shorthand notation for $[0, D]^k$. We are interested in the regimes for which the Gibbs model provides a higher cache hit probability than independent placement, i.e., $P_{\text{Miss,G}}(m, k) \leq (1 - p_c(m))^k$, where $p_c(m)$ is the placement probability for index m .

Remark 1. Necessary conditions on $\theta_m(\cdot)$ to make the cache hit optimization problem convex. Consider the cache hit maximization problem in (20), which is equivalent to the minimization of the cache miss. If the average cache miss function is a convex function and the constraint is also convex, then if a local minimum exists for (20), it is also the global minimum. The average cache miss function $\sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N} = k) P_{\text{Miss,G}}(m, k)$ is convex in β_m if the function $P_{\text{Miss,G}}(m, k) = \int \cdots \int_{V^k} f_m(x_1, \dots, x_k) dx_1 \dots dx_k$ is convex in β_m .

Remark 2. If $f(x, y)$ is convex for each $y \in A$, and $w(y) \geq 0$ for each $y \in A$, then the function g defined as $g(x) = \int_A w(y) f(x, y) dy$ is convex in x , provided the integral exists [34, Ch. 3.2].

Remark 3. A twice differentiable function h with convex domain $\text{dom } h$ is convex if and only if, $\Delta^2 h(x) \succeq 0$ for all $x \in \text{dom } h$.

From Remarks 1 and 3, $f(x) = \exp(-\theta(x))$ is convex if and only if $\frac{\partial^2 f(x)}{\partial x^2} = [(\theta'(x))^2 - \theta''(x)] \exp(-\theta(x)) \geq 0$. Hence, $\theta(x)$ should be chosen such that $(\theta'(x))^2 - \theta''(x) \geq 0$. From Remark 2, as long as $f_m(x_1, \dots, x_k)$ is convex in for all $\{x_1, \dots, x_k\} \in V^k$, $P_{\text{Miss,G}}(m, k)$ is convex.

Remark 4. Product of convex and nondecreasing (nonincreasing) and positive functions is convex.

Proposition 3. From Remark 4, the function $f_m(x_1, \dots, x_k) = \exp(-\beta_m \sum \sum_{1 \leq i < j \leq k} \theta(\|x_i - x_j\|)) = \prod \prod_{1 \leq i < j \leq k} \exp(-\beta_m \theta(\|x_i - x_j\|))$ is convex if $\theta(x)$ is chosen according to Remark 3.

Remark 5. Requirements on the potential function for optimal content placement. From the theory of statistical mechanics, distributions of the form given in (21) arise when the point process is constrained to have fixed mean energy

$$\bar{E}_k = \int_{D^k} E(x_1, \dots, x_k) f(x_1, \dots, x_k) dx_1 \dots dx_k. \quad (22)$$

Proposition 4. Increasing the pair potential function $\theta(\cdot)$ improves the cache hit probability.

Proof: From (17), $E(x_1, \dots, x_k)$ increases with $\theta(\cdot)$ for a set of points $\{x_1, \dots, x_k\}$. Due to the constant mean energy as in (22), $E(x_1, \dots, x_k)$ and $f(x_1, \dots, x_k)$ should be negatively dependent.

The miss probability $\int \cdots \int_{V^k} f_m(x_1, \dots, x_k) dx_1 \dots dx_k$ becomes small if $f_m(x_1, \dots, x_k)$ takes smaller values, or equivalently $E(x_1, \dots, x_k)$ takes larger values, for the k dimensional region $[0, D]^k \setminus [0, R_{D2D}]^k$. Hence, increasing $\theta(\cdot)$ improves the cache hit probability. However, from (22), it is not possible to increase $E(x_1, \dots, x_k)$ unboundedly. Therefore, $\theta(\cdot)$ may not be increased arbitrarily. However, the best tradeoff between $\theta(\cdot)$ and $f_m(x_1, \dots, x_k)$ can be found to maximize the cache hit probability. ■

Proposition 5. *Given the pairwise potential $\theta(r)$ is monotonic, it should be decreasing with r to achieve a higher cache hit probability.*

Proof: Let $\theta(r)$ be a monotonic and positive potential that is decreasing in r . Under a total energy constraint, for small r , where $\theta(r)$ or equivalently $E(x_1, \dots, x_k)$ is high, $f_m(x_1, \dots, x_k)$ is small. On the other hand, for large r , where $\theta(r)$ or equivalently $E(x_1, \dots, x_k)$ is small, the distribution $f_m(x_1, \dots, x_k)$ takes larger values. The distribution $f_m(x_1, \dots, x_k)$ is given by

$$f_m(x_1, \dots, x_k) = \exp(-\beta_m E(x_1, \dots, x_k)) / Z_{m,k}, \quad (23)$$

where $Z_{m,k} = \int \cdots \int_{D^k} \exp(-\beta_m E(x_1, \dots, x_k)) dx_1 \dots dx_k$. We require $\int \cdots \int_{V^k} f_m(x_1, \dots, x_k) dx_1 \dots dx_k$ to be small to achieve higher hit probability. Therefore, we infer that decreasing potential $\theta(\cdot)$ is required for content placement to yield higher cache hit probabilities. ■

Since the original formulation (20) is not tractable, we next consider an approximate solution.

An approximate solution for Gibbs placement. An approximate solution for the Gibbs optimization problem in (20) can be found by using a similar approximation as in (19) of [33]. Therefore, we have the relation $Z_{m,k} \approx (\pi D^2)^k \left(\frac{2}{D^2} \int_0^D \exp(-\beta_m \theta(r)) r dr \right)^{k(k-1)/2}$, and the probability that k caches cover a receiver, and none has file m is approximated as

$$P_{\text{Miss,G}}(m, k) \approx \frac{(\pi(D^2 - R_{D2D}^2))^k \left(\frac{2}{(D^2 - R_{D2D}^2)} \int_{R_{D2D}}^D \exp(-\beta_m \theta(r)) r dr \right)^{k(k-1)/2}}{(\pi D^2)^k \left(\frac{2}{D^2} \int_0^D \exp(-\beta_m \theta(r)) r dr \right)^{k(k-1)/2}}. \quad (24)$$

The function $f_m(x_1, \dots, x_{N_m+1})$ is also approximated as follows:

$$f_m(x_1, \dots, x_{N_m+1}) \approx \frac{\exp\left(-\beta_m \sum_{1 \leq i < j \leq N_m+1} \theta(\|x_i - x_j\|)\right)}{(\pi D^2)^{N_m+1} \left(\frac{2}{D^2} \int_0^D \exp(-\beta_m \theta(r)) r dr \right)^{(N_m+1)N_m/2}}. \quad (25)$$

We approximate the LHS of the constraint of the cache hit maximization problem in (20) as follows:

$$\sum_{m=1}^M \mathbb{E}_{\mathcal{N}} [f_m(x_1, \dots, x_{\mathcal{N}+1})] \approx \sum_{m=1}^M \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N} = k) \frac{\mathbb{E} \left[e^{-\beta_m \sum_{1 \leq i < j \leq k+1} \theta(\|x_i - x_j\|)} \right]}{(\pi D^2)^{k+1} \left(\frac{2}{D^2} \int_0^D e^{-\beta_m \theta(r)} r dr \right)^{\frac{(k+1)k}{2}}}, \quad (26)$$

where

$$\begin{aligned} \mathbb{E} \left[e^{-\beta_m \sum_{1 \leq i < j \leq k+1} \theta(\|x_i - x_j\|)} \right] &\stackrel{(a)}{=} \prod_{1 \leq i < j \leq k+1} \mathbb{E} [e^{-\beta_m \theta(\|x_i - x_j\|)}] \\ &\stackrel{(b)}{=} \left(\frac{1}{D^2} \int_0^D \int_0^D e^{-\beta_m \theta(\|x_i - x_j\|)} dx_i dx_j \right)^{\frac{(k+1)k}{2}} \\ &\stackrel{(c)}{=} \left(\frac{1}{2\pi} \int_0^{2\pi} \frac{1}{D^2} \int_0^D \int_0^D e^{-\beta_m \theta(\sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos(\phi)})} r_1 r_2 dr_1 dr_2 d\phi \right)^{\frac{(k+1)k}{2}}, \end{aligned}$$

where (a) follows from the independence, (b) follows from that given the realization of the Poisson distributed random variable \mathcal{N} in the bounded interval $[0, D]$, x_i 's are uniformly distributed, and (c) follows from algebraic manipulation and converting Cartesian to polar coordinates.

The hit probability can be approximated as $P_{\text{Hit,G}} = 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N} = k) P_{\text{Miss,G}}(m, k)$, using (24) and optimizing the β_m values using (26) and according to the popularity profile $p_r(m)$.

With the approximation in (24), it is clear that we do not require any positivity condition on the pair potential function $\theta(\cdot)$ as shifting the potential function by a constant factor, i.e., $\theta'(r) = \theta(r) - a < 0$, would yield the same cache miss probability. However, in this paper, due to tractability reasons, we restrict ourselves to positive potentials, which yields repulsion, i.e., negatively correlated placement.

Next we consider some special cases of GPPs that yield repulsion. Negatively correlated spatial placement corresponds to a distance-dependent thinning of the transmitter process so that neighboring users are less likely to have matching contents.

VI. REPULSIVE CACHE PLACEMENT

We discussed GPPs modeling the pairwise interactions between users and showed that repulsion can provide a higher cache hit in Sect. V. The focus of this section is on repulsive, i.e., negatively dependent, placement strategies. We start with the definition of negative dependence.

Definition 4. Random variables Y_1, \dots, Y_k , $k \geq 2$, are said to be negatively dependent, if for any numbers $y_1, \dots, y_k \in \mathbb{R}$, we have that [35]

$$\mathbb{P}\left(\bigcap_{m=1}^k Y_m \leq y_m\right) \leq \prod_{m=1}^k \mathbb{P}(Y_m \leq y_m), \quad \mathbb{P}\left(\bigcap_{m=1}^k Y_m > y_m\right) \leq \prod_{m=1}^k \mathbb{P}(Y_m > y_m). \quad (27)$$

Proposition 6. Negatively dependent placement performs better than independent placement in terms of the average cache hit probability.

Proof: For a negatively dependent identical content placement, we can infer that $P_{\text{Miss},N}(m, k) \stackrel{(a)}{\leq} \prod_{m=1}^k \mathbb{P}(Y_m = 0) \stackrel{(b)}{=} \mathbb{P}(Y_m = 0)^k$, where (a) comes from Defn. 4, and (b) is from identical content placement assumption. Hence, for a negatively dependent content placement strategy, the hit probability satisfies $P_{\text{Hit},N} = 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N} = k) P_{\text{Miss},N}(m, k) \stackrel{(a)}{\geq} 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N} = k) \mathbb{P}(Y_m = 0)^k$, where the RHS of (a) is the hit probability for independent placement for $p_c(m) = 1 - \mathbb{P}(Y_m = 0)$. \blacksquare

We ask the following question: Given the coverage number k and file m , how large cache hit rates can we achieve, i.e., how small can $P_{\text{Miss},N}(m, k) \leq \mathbb{P}(Y_m = 0)^k$ get for a spatial content placement setting, or what is the best negatively dependent content placement strategy? To answer that, we next consider negatively dependent cache placement strategies inspired from GPPs.

Permutation distribution. Random variables having the permutation distribution are negatively dependent [36]. The random variables Y_1, \dots, Y_n have the permutation distribution on $\{0, 1\}$, if they take values in $\{0, 1\}$, i.e., $y_m \in \{0, 1\}$, $m \in \{1, \dots, n\}$ and for every permutation $y_{r(1)}, \dots, y_{r(n)}$ of y_1, \dots, y_n , it is satisfied that $\mathbb{P}(Y_1 = y_{r(1)}, \dots, Y_n = y_{r(n)}) = \frac{1}{n!}$ [37].

We next consider a special case of permutation invariant schemes.

Fermi-Dirac model. The Fermi-Dirac (F-D) model, a special case of GPPs, is a statistical physics model to describe a distribution of particles over single-particle energy states in systems consisting of many identical particles, where a many particle system is described in terms of single particle energy states. Using the F-D distribution [38] for a system of identical particles, the average number of particles in a single-particle state ϵ is given by a sigmoid function

$$F(\epsilon) = \frac{1}{\exp((\epsilon - \mu)/k_B T) + 1} = \frac{1}{\exp(\beta(\epsilon - \mu)/k_B) + 1}, \quad (28)$$

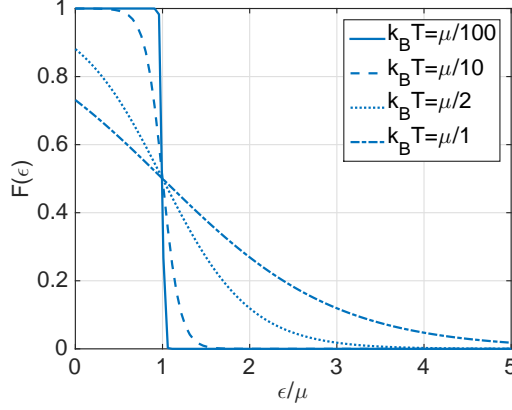


Fig. 3: Fermi-Dirac distribution.

which is the Fermi function, where k_B is Boltzmann's constant, $T = \beta^{-1}$ is the absolute temperature, which determines the shape of the distribution, ϵ is the energy of the single-particle state, and μ is the total chemical potential. At zero temperature, μ is equal to the Fermi energy plus the potential energy per particle [39]. The F-D distribution is only valid if the number of particles in the system is large enough so that adding one more particle to the system has negligible effect on μ [38]. Since the F-D distribution is derived using the Pauli exclusion principle⁵, it is required that $0 < F(\epsilon) < 1$. The Fermi function $F(\epsilon)$, i.e., the F-D distribution, is illustrated in Fig. 3. As the inverse temperature parameter β increases, $F(\epsilon)$ also increases for $\epsilon < \mu$, and decreases for $\epsilon > \mu$.

The average number of particles with energy ϵ per unit volume is $\bar{N}(\epsilon) = g(\epsilon)F(\epsilon)$, where $g(\epsilon)$ is called the degeneracy [39], i.e., the number of states with energy ϵ per unit volume [41]. When $g(\epsilon) \geq 2$, it is possible that $\bar{N}(\epsilon) > 1$ since there is more than one state that can be occupied by particles with the same energy ϵ . The F-D distribution can be extended to the multi-particle case.

Exploiting the F-D model for Caching. The F-D model is equivalent to having a bin-ball (cache-content) problem, in which a cache contains at most one content of the same type, and each distribution of balls among the bins is equally likely to occur. Given the indicator variables Y_i 's for the F-D model, the joint distribution of (Y_1, \dots, Y_n) in the F-D model for n random variables is given as $\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n) = \binom{n}{t_n}^{-1}$ for a given realization $y_i \in \{0, 1\}$,

⁵The Pauli exclusion principle states that two identical fermions cannot occupy the same quantum state simultaneously [40].

$i \in [n]$, with $\sum_i y_i = t_n$ [37]. Using the F-D model for the multi file model with $Y_{m,i}$'s for file m , $P_{\text{Miss},F}(m, k) = \mathbb{P}(\bigcap_{i=1}^k \{Y_{m,i} = 0\}) = \mathbb{P}(t_{m,k} = 0)$. However, the challenge to optimize $t_{m,k}$ for all different file types. The F-D model will help determine the number of replications for each content type. The parameter T should be adjusted based on the skewness of the demand distribution to determine the shape of the caching distribution.

Exploiting the F-D model, $F_m(\epsilon)$ gives the probability that file m is placed in i^{th} cache as

$$F_m(\epsilon) = \mathbb{P}(Y_{m,i} = 1) = \frac{1}{\exp(\beta_m(\epsilon - \mu)/k_B) + 1}. \quad (29)$$

For given parameters ϵ, μ , the cache placement condition to be satisfied is $\sum_{m=1}^M F_m(\epsilon) \leq N$.

Given the degeneracy $g_m(\epsilon)$ of each file⁶, i.e., the number of users that contain file m for a given ϵ , the joint cache placement distribution is

$$\mathbb{P}(Y_{m,1} = y_{m,1}, \dots, Y_{m,k} = y_{m,k}) = \binom{k}{g_m(\epsilon)}^{-1}, \quad \sum_{i=1}^k y_{m,i} = g_m(\epsilon). \quad (30)$$

The inverse temperature parameters β_m vary with respect to the file popularities. For high popular files, β_m becomes large, yielding a sharper $F_m(\epsilon)$, i.e., more deterministic placement, and for low popular files, β_m is small, yielding a more spread out placement distribution. Note that there are also some special cases of GPPs, with various ranges of β , such as $\beta_m \in [0, \infty)$ as we will detail in this section. In Sect. VII, we show that the parameter β in some cases does not affect the optimal placement distribution, e.g., in the case of hard-core placement with iid marking, popular files correspond to small R_m , and less popular files have larger exclusion radius, where the potential function satisfies $\theta \in \{0, \infty\}$ and the variation of β does not affect the states.

The cache hit probability formulation for the F-D distribution-inspired caching model is given as

$$\begin{aligned} \max_{\beta_m} \quad & P_{\text{Hit},F} \\ \text{s.t.} \quad & \sum_{m=1}^M \mathbb{E}_{\mathcal{N}} [f_m(x_1, \dots, x_{N+1})] \leq N, \end{aligned} \quad (31)$$

⁶Although the optimal values of $g_m(\epsilon) \in \mathbb{Z}^+$ are nontrivial to assign, the probabilistic placement policy in [12] can be exploited that yields $g_m(\epsilon) \approx p_c^*(m)k$, where $p_c^*(m)$ is the optimal cache placement distribution for independent placement.

where $P_{\text{Hit},F} = 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N} = k) P_{\text{Miss},F}(m, k)$ is the cache hit probability, where $\mathbb{P}(\mathcal{N} = k) = \exp(-\lambda_t \pi R_{\text{D2D}}^2) \frac{(\lambda_t \pi R_{\text{D2D}}^2)^k}{k!}$, and $P_{\text{Miss},F}(m, k) = \sum_{s \in \mathcal{S}_f} \frac{1}{Z_{m,k}} \exp(-\beta_m(\epsilon_s - \mu k))$, where \mathcal{S}_f is the set of states that yields failure and $Z_{m,k} = \sum_{s \in \mathcal{S}} \exp(-\beta_m(\epsilon_s - \mu k))$, where \mathcal{S} is the set of all possible states of the many-file system. Considering a continuum of energies and given an energy threshold $E_{T,m}$ for file m , we can rewrite $P_{\text{Miss},F}(m, k)$ as

$$P_{\text{Miss},F}(m, k) = \frac{\int_{E_{T,m}}^{\infty} \exp(-\beta_m(\epsilon - \mu k)) d\epsilon}{\int_0^{\infty} \exp(-\beta_m(\epsilon - \mu k)) d\epsilon} = \exp(-\beta_m E_{T,m}). \quad (32)$$

Thus, $P_{\text{Hit},F} = 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N} = k) \exp(-\beta_m E_{T,m}) = 1 - \sum_{m=1}^M p_r(m) \exp(-\beta_m E_{T,m})$.

The constraint of (31) is equivalent to

$$\sum_{m=1}^M \mathbb{E}_{\mathcal{N}_m} \left[\frac{\int_0^{E_{T,m}} \exp(-\beta_m(\epsilon - \mu(\mathcal{N}_m + 1))) d\epsilon}{\int_0^{\infty} \exp(-\beta_m(\epsilon - \mu(\mathcal{N}_m + 1))) d\epsilon} \right] = \sum_{m=1}^M (1 - \exp(-\beta_m E_{T,m})) \leq N, \quad (33)$$

where note that as the threshold $E_{T,m}$ increases, the cache constraint can be violated.

Thus, the formulation in (31) is equivalent to

$$\begin{aligned} \min_{\beta_m} \quad & \sum_{m=1}^M p_r(m) \exp(-\beta_m E_{T,m}) \\ \text{s.t.} \quad & M - N \leq \sum_{m=1}^M \exp(-\beta_m E_{T,m}), \end{aligned} \quad (34)$$

where it is easy to observe that the optimal solution is $\beta_m = \infty$ for $m \in \{1, \dots, N\}$ and $\beta_m = 0$ for $m \in \{N+1, \dots, M\}$, leads to the deterministic placement of the most popular files for $\epsilon < \mu$ (See (29) and Fig. 3.).

If instead we consider a single parameter β for all files and given different energy thresholds $E_{T,m}$'s for different files, which are determined based on the same F-D distribution where $E_{T,m}$ increases with popularity, i.e., decreases with m , the formulation in (34) can be easily solved, where the optimal value of β satisfies $M - N = \sum_{m=1}^M \exp(-\beta^* E_{T,m})$ because if $\beta < \beta^*$, the constraint is satisfied with inequality, but the miss probability increases, and when $\beta > \beta^*$, the constraint is violated.

Another way to solve the caching problem is to fix β and consider a continuum of energies ϵ , and vary ϵ to optimize the cache hit probability for a given popularity profile $p_r(m)$. Denoting

the F-D distribution by $F(\epsilon)$, for given μ, β and a continuum of energies ϵ , the cache placement condition to be satisfied is $\int_0^\infty F(\epsilon)d\epsilon = \int_0^\infty \frac{1}{\exp(\beta(\epsilon-\mu)/k_B)+1}d\epsilon \leq N$, which is left as future work.

In addition to F-D models, general GPPs can be studied for different pairwise interactions determined by the potential function $\theta(\cdot)$, to observe the cache hit probability performance limits of spatial content caching through the general optimization formulation given in (20).

VII. MATÉRN HARD-CORE MODEL-BASED CONTENT PLACEMENT

We next consider the hard-core regime for the GPPs, which provides useful insights for the development of spatial content placement for the regime relevant to D2D communications. We provide two different content placement models both inspired from the Matérn hard-core (type II) model, where the Gibbs pair potential takes a simple form that makes our analysis tractable.

A. Matérn Hard-Core Placement Model I

Matérn's hard-core (MHC) model is a special case of the permutation distribution in which we pick a subset of transmitters based on some exclusion that yields the negative dependence among the nodes.

We propose a content placement approach exploiting the spatial properties of MHC model type II, which we call MHC-A. This type of MHC model is constructed from the underlying PPP modeling the locations of the caches by removing certain points depending on the positions of the neighboring points and additional marks attached to the points. Each transmitter of the BM V_{BM} is assigned a uniformly distributed mark $U[0, 1]$. A node $x \in \tilde{\Phi}$ is selected if it has the lowest mark among all the points in $B_x(R_{\text{D2D}})$. A realization of the MHC point process is illustrated in Fig. 4. The proposed placement model is slightly different. Instead, for each file type, there is a distinct exclusion radius (r_m for file m).

We optimize the exclusion radii to maximize the total hit probability. The exclusion radius of a particular file r_m depends on the file popularity in the network, transmitter density and the cache size and satisfies $r_m < R_{\text{D2D}}$. Otherwise, once r_m exceeds R_{D2D} , as holes would start to open up in the coverage for that content, the hit probability for file m would suffer. We consider the following cases: (i) if the file is extremely popular, then many transmitters should simultaneously cache the file, yielding a small exclusion radius, and (ii) if the file is not very popular, then fewer

(or zero) transmitters would be sufficient for caching the file, yielding a larger exclusion radius. Therefore, intuitively, we might expect the exclusion radius to decrease with increasing file popularity. However, our analysis shows that the exclusion radius is positively correlated with the file popularity, i.e., the most popular files are stored in a few caches with higher marginal probabilities unlike the files with low popularity that are stored with lower marginals, but with smaller exclusion radius.

Given the exclusion radius of the MHC-A model, a file should be placed at only one cache within a circular region. Hence, the caching probability of file m at a typical transmitter is

$$p_{\text{cache}}(m) \stackrel{(a)}{=} \mathbb{E}\left[\frac{1}{1 + C_m}\right] = \frac{1 - \exp(-\bar{C}_m)}{\bar{C}_m}, \quad (35)$$

where $C_m \sim \text{Poisson}(\bar{C}_m)$ is number of neighboring transmitters in a circular region of radius r_m with mean $\bar{C}_m = \lambda_t \pi r_m^2$, and (a) follows from the fact that the caching probability of a typical transmitter equals the probability that the node qualifies and gets the minimum mark value in its neighborhood.

Let \tilde{C}_m be the number of transmitters containing file m within a circular region of radius r_m . Since only one transmitter is allowed to contain a file within the exclusion radius, $\tilde{C}_m \in \{0, 1\}$. Given the MHC-A model, a transmitter having file m in a region of size πr_m^2 exists with probability

$$\mathbb{P}(\tilde{C}_m = 1) = 1 - \exp(-\bar{C}_m). \quad (36)$$

Hence, $\mathbb{E}[\tilde{C}_m] = \lambda_{\text{MHC}}(m) \pi r_m^2 = 1 - \exp(-\bar{C}_m)$ [24, Ch. 2.1], where $\lambda_{\text{MHC}}(m)$ is the density of the MHC-A model. The maximum hit probability for the MHC-A model is given by the solution of:

$$\begin{aligned} \max_{p_{\text{cache}}(m)} \quad & P_{\text{Hit},M} = \sum_{m=1}^M p_r(m) \mathbb{P}(\tilde{C}_m = 1) \\ \text{s.t.} \quad & \sum_{m=1}^M p_{\text{cache}}(m) \leq N. \end{aligned} \quad (37)$$

We define the Lagrangian to find the solution as follows:

$$\mathcal{M}(\zeta) = \sum_{m=1}^M p_r(m) (1 - e^{-\bar{C}_m}) + \zeta \left(\sum_{m=1}^M \frac{1 - e^{-\bar{C}_m}}{\bar{C}_m} - N \right). \quad (38)$$

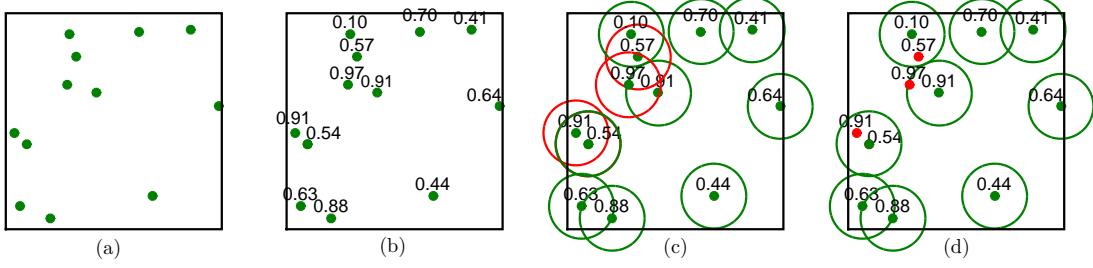


Fig. 4: MHC point process realization: (a) Begin with a PPP. (b) Associate a mark $\sim U[0, 1]$ to each point independently. (c) A node x is selected if it has the lowest mark among all the points in $B_x(R_{D2D})$. (d) Set of selected points.

Evaluating this at $\left. \frac{d\mathcal{M}(\zeta)}{d\bar{C}_m} \right|_{\zeta=\zeta^*} = 0$, we obtain the simplified relation $\zeta^* = h_m(\bar{C}_m^*) = \frac{p_r(m)(\bar{C}_m^*)^2}{(\exp(\bar{C}_m^*) - \bar{C}_m^* - 1)}$, where the optimal solution ζ^* depends on the cache size N . Note that ζ^* is decreasing in \bar{C}_m , $\lim_{\bar{C}_m \rightarrow 0} \zeta^* = 2p_r(m)$ and $\lim_{\bar{C}_m \rightarrow \infty} \zeta^* = 0$. We determine the optimal value of \bar{C}_m as

$$\bar{C}_m^* = \begin{cases} 0 & \text{if } \zeta^* \geq 2p_r(m), \\ h_m^{-1}(\zeta^*) & \text{if } \zeta^* < 2p_r(m). \end{cases} \quad (39)$$

For very unpopular files with small $p_r(m)$, ζ^* satisfies $\zeta^* > 2p_r(m)$ and hence, $\bar{C}_m^* = 0$. As the file popularity increases, $p_r(m)$ will be higher and ζ^* satisfies the relation $\zeta^* \leq 2p_r(m)$. Hence \bar{C}_m^* increases with popularity and satisfies the relation $h_m^{-1}(\zeta^*)$. Thus, the average number of transmitters within the exclusion region, i.e., \bar{C}_m^* , is increasing by increasing the file popularity, and the exclusion radius for files with high popularity should be higher, which yields lower $p_{\text{cache}}(\cdot)$ for popular files from (35). If the demand distribution is uniform over the network, then each file has the same caching probability, i.e., $p_{\text{cache}}(m)$ is the same for all m , yielding the same r_m for all m , which is intuitive. When the demand distribution is skewed towards the more popular files, then the required condition is $r_m \propto 1/\sqrt{\lambda_{\text{MHC}}(m)}$, i.e., more popular files will end up being stored in fewer locations.

Combining (36) with the fact that \bar{C}_m^* increases with the file popularity, we have $\mathbb{P}(\tilde{C}_1 = 1) > \mathbb{P}(\tilde{C}_2 = 1) > \dots > \mathbb{P}(\tilde{C}_M = 1)$, which implies that the probability of having one transmitter having a file increases with its popularity. This is intuitive because as the exclusion radius for a particular file becomes higher, the volume fraction grows, and the probability of having a transmitter caching that file increases. For the case when the exclusion radius is smaller, a larger

fraction of the transmitters are prevented from caching and hence finding a transmitter for a particular file within the exclusion region becomes smaller. Popular files are guaranteed to be available over a larger geographic area. On the contrary, files with low popularity are available with low probability.

2nd Order Product Density. Consider a circular region \mathcal{D} of radius D with $D \gg r_1 > \dots > r_M$ and let $\bar{C}_D = \lambda_t \pi D^2$ be the average number of transmitters within \mathcal{D} . Due to the limited storage capacity of the caches, the mean total number of files that can be cached in region \mathcal{D} is upper bounded by $N\bar{C}_D$. To determine the average number of users containing a desired file type in region \mathcal{D} , we use the second order product density of the MHC process. The second-order product density of the MHC process $\tilde{\Phi}$ is defined as the joint probability that there are two points of $\tilde{\Phi}$ at two specified locations x and y in the infinitesimal volumes dx and dy [42], and is given by [43]

$$\rho_m^{(2)}(r) = \begin{cases} \lambda_{\text{MHC}}^2(m), & r \geq 2r_m \\ \frac{2V_{r_m}(r)[1-\exp(-\lambda_t \pi r_m^2)]}{\pi r_m^2 V_{r_m}(r)[V_{r_m}(r)-\pi r_m^2]} - \frac{2\pi r_m^2 [1-\exp(-\lambda_t V_{r_m}(r))]}{\pi r_m^2 V_{r_m}(r)[V_{r_m}(r)-\pi r_m^2]}, & 2r_m > r > r_m \\ 0, & r \leq r_m \end{cases}, \quad (40)$$

where $V_{r_m}(r) = 2\pi r_m^2 - 2r_m^2 \cos^{-1}\left(\frac{r}{2r_m}\right) + r\sqrt{r_m^2 - \frac{r^2}{4}}$ denotes the area of the union of two circles having radius r_m and separated by distance r .

Active Transmitters. For a stationary point process $\tilde{\Phi}$, using Campbell's theorem [24, Ch. 1.4], we can deduce that the average number of transmitters (conditioned on there being a point at the origin but not counting that point) contained in the ball $B_0(R_{\text{D2D}})$ is given by

$$\mathbb{E}^{\text{!o}} \left[\sum_{x \in \tilde{\Phi}} 1(x \in B_0(R_{\text{D2D}})) \right] = \lambda_t^{-1} \int_{B_0(R_{\text{D2D}})} \rho^{(2)}(x) dx. \quad (41)$$

Using (41), we can deduce the average number of simultaneously active transmitters. If $r_m < R_{\text{D2D}}$, then (41) yields a positive number. Therefore, conditioned on there being a point at the origin, there is also at least a point within distance R_{D2D} with positive probability. In this case, if (41) yields a number which is greater than or equal to 1, then there is a transmitter covering the user. If (41) is less than 1, it is equal to the probability that there is a user covering the typical receiver. However, when $r_m \geq R_{\text{D2D}}$, then (41) yields 0. Therefore, the probability that the user is covered is determined by the probability that there exists a transmitter at the origin

as determined by (36).

Incorporating the finiteness of R_{D2D} into the MHC-A model, (36) becomes

$$\mathbb{P}(\tilde{C}_m = 1) = \begin{cases} \min\{2\pi\lambda_t^{-1} \int_{r_m}^{R_{D2D}} \rho_m^{(2)}(r)r dr, 1\} & \text{if } r_m < R_{D2D}, \\ 1 - \exp(-\bar{C}_m), & \text{if } r_m \geq R_{D2D}, \end{cases} \quad (42)$$

which will determine the cache hit probability.

From (41), the average number of users caching file m in $B_0(D)$ denoting the region \mathcal{D} is given by

$$N_m = \lambda_t^{-1} \int_{B_0(D)} \rho_m^{(2)}(x) dx = 2\pi\lambda_t^{-1} \int_{r_m}^D \rho_m^{(2)}(r)r dr. \quad (43)$$

Ideally, when the MHC placement strategy is applied, all the files need to be placed at the caches in a way that all the cache slots are occupied. Due to the random assignment of the marks in the MHC approach, it is not guaranteed that all the caches are full, which causes underutilization of the caches. Thus, $\sum_{m=1}^M N_m \leq N\bar{C}_D$. The thinning leading to the MHC process can be refined such that higher intensities λ_{MHC} are possible [28, Ch. 5.4], at the price of more complicated algorithms [44] and [45].

As the storage size of the users drops, the exclusion region should increase to bring more spatial diversity into the model. Using the cache storage constraint in (37), as N drops, a typical receiver won't be able to find its requested files and $\lim_{N \rightarrow 0} r_m = \infty$, which increases the volume fraction, i.e., increases the caching probability. When N increases sufficiently, the transmitter candidates of the typical receiver will have any of the requested files and $\lim_{N \rightarrow \infty} r_m = 0$, and because it is redundant to cache the files at all the transmitters, the exclusion radius should be made smaller to decrease the volume fraction and the caching probability. Thus, N and r_m have an inverse relationship.

Remark 6. *The MHC-A model is a special case of the Gibbs model discussed in Sect. V, also known as Gibbs hard-core process, and its pair potential function satisfies $\theta_m(r) = \infty$ when $r \leq r_m$, and $\theta_m(r) = 0$ if $r > r_m$ for a given hard-core exclusion radius r_m for file m , assuming $R_{D2D} \leq r_m$, and*

$$f_m(x_1, \dots, x_k) = \exp\left(-\beta \sum_{1 \leq i < j \leq k} \theta_m(\|x_i - x_j\|)\right) / Z_{m,k} = 1 / Z_{m,k}$$

is the distribution of the point process, i.e., f is uniformly distributed.

Another way to obtain the uniform distribution is to let the temperature be $T = \infty$ so that $\beta = 0$ and f is uniform no matter what the trend of the potential function $\theta(r)$. Hence, the constraint of the optimization formulation for GPP model in (20) becomes

$$\sum_{m=1}^M \mathbb{E}_{\mathcal{N}_m} [f_m(x_1, \dots, x_{\mathcal{N}_m+1})] = \sum_{m=1}^M \mathbb{E} \left[\frac{1}{\mathcal{N}_m + 1} \right] \leq N,$$

which is the cache placement constraint of the MHC-A model, where $\mathcal{N}_m \sim \text{Poisson}(\lambda_t \pi r_m^2)$.

The cache miss probability is given by

$$\begin{aligned} P_{\text{Miss,G}}(m, k) &\stackrel{(a)}{=} \frac{\int_{\text{R}_{\text{D2D}}}^D \cdots \int \exp \left(-\beta \sum_{1 \leq i < j \leq k} \theta_m(\|x_i - x_j\|) \right) dx_1 \dots dx_k}{\int_0^D \cdots \int \exp \left(-\beta \sum_{1 \leq i < j \leq k} \theta_m(\|x_i - x_j\|) \right) dx_1 \dots dx_k} \\ &\stackrel{(b)}{=} \frac{\left(\int_{r_m}^D dx \right)^k}{\left(\int_0^D dx \right)^k} \stackrel{(c)}{=} \left(1 - \frac{r_m^2}{D^2} \right)^k \end{aligned} \quad (44)$$

where (a) follows from the definition of cache miss region, (b) follows from that $\theta_m(r) = 0$ for $r > r_m$ and $\theta_m(r) = \infty$ for $r \leq r_m$, and hence $\theta_m(\text{R}_{\text{D2D}}) = \infty$, and (c) follows from converting the integral into polar coordinates. Using (44), and that the number of users in the region with radius D is Poisson with $\mathbb{P}(\mathcal{N} = k) = e^{-\lambda_t \pi D^2} \frac{(\lambda_t \pi D^2)^k}{k!}$, the cache hit probability is

$$\begin{aligned} P_{\text{Hit,M}} &= 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} e^{-\lambda_t \pi D^2} \frac{(\lambda_t \pi D^2)^k}{k!} P_{\text{Miss,G}}(m, k) \\ &= \sum_{m=1}^M p_r(m) (1 - \exp(-\lambda_t \pi r_m^2)), \end{aligned} \quad (45)$$

which is the same as the solution of (37) given the r_m values are optimized. The case where $\text{R}_{\text{D2D}} > r_m$ and $\theta_m(\text{R}_{\text{D2D}}) = 0$, the miss probability in (44) is more difficult to evaluate because $f_m(x_1, \dots, x_k)$ is no longer uniformly distributed⁷. In that case, the miss probability can be approximated using (24).

⁷Pairwise correlations between the points separated by $r > r_m$ are modeled using the second-order product density $-\rho_m^{(2)}(r)$ for file m - of the MHC process as previously discussed.

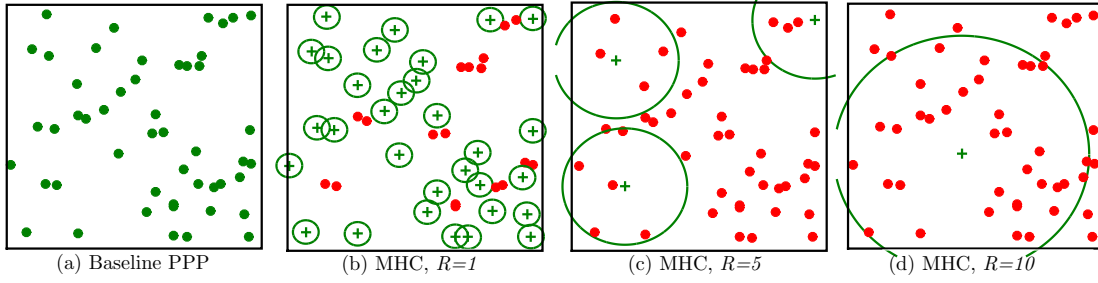


Fig. 5: MHC versus the exclusion radii. Selected nodes are denoted by plus sign.

We next detail a different MHC-based model called MHC-B without the need to solve a cache hit maximization problem and provide sufficient conditions for achieving high cache hit probability.

B. Matérn Hard-Core Placement Model II

In this section, we propose a new MHC-based placement model called MHC-B. Unlike the MHC-A model in Sect. VII-A, where we maximize the average cache hit probability given the finite cache storage constraint, we optimize the exclusion radii and provide sufficient conditions so that the MHC-B model is at least as good as the independent placement model of Sect. III-III-B.

The proposed content placement model is slightly different from the MHC point process transmission model with fixed radius. Instead, for each file type, there exists a different exclusion radius. For each file type, a circular exclusion region is created around each active transmitter to prevent all the transmitters located in a circular region from caching a particular content simultaneously. The exclusion radii is determined by the file popularity, which will be detailed next.

The critical exclusion radius should be inversely proportional to the popularity of the requests, which is mainly determined by the skewness parameter γ_r . As γ_r increases, the distribution becomes more skewed and higher variability should be observed in the exclusion radii of different files.

Optimizing the marginal distribution for content caching is not sufficient to optimize the joint performance of the caching. Therefore, in this section, we consider a correlated content caching model over space that improves the performance of caching with the same marginal caching

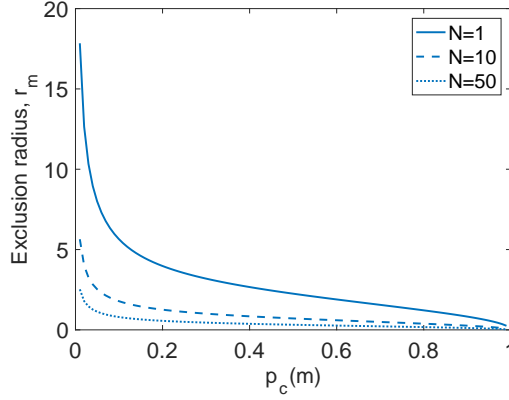


Fig. 6: Characterization of optimal exclusion radii for $N = [1, 10, 50]$, $M = 100$, $\lambda_t = .1$, $R_{D2D} = 1$.

probabilities. Different from the model in Sect. VII, we determine the exclusion radii based on the independent placement model, i.e., on average the fraction of the users containing a file is equal to its optimal placement probability in the independent placement model.

In Fig. 5, we illustrate the trend of the MHC process thinned from the baseline PPP for different exclusion radii. As the exclusion radius R increases, the intensity λ_{MHC} of MHC-B process decreases.

Lemma 2. *To achieve a better average cache hit probability performance than the independent placement, the exclusion radius for content m for the MHC-B model should be selected as*

$$r_m \leq \sqrt{\frac{1}{N\lambda_t\pi} W\left(-\frac{\exp(-1/p_c^*(m))}{p_c^*(m)}\right) + \frac{1}{N\lambda_t\pi p_c^*(m)}},$$

where $p_c^*(\cdot)$ is the placement probability for independent placement and W is the Lambert function.

In Fig. 6, we show the trend of the exclusion radius r_m with respect to the caching pmf $p_c(m)$. As we expect, the exclusion radius r_m decays with the popularity and the cache size N . Furthermore, from (36), $r_m > R_{D2D} \sqrt{p_c^*(m)}$ is a sufficient condition for MHC-B to be better than the independent placement model in terms of hit probability, as long as $r_m \leq R_{D2D}$ is satisfied. However, for files with very low popularity, r_m tends to be very high, and the $r_m \leq R_{D2D}$ condition might be violated.

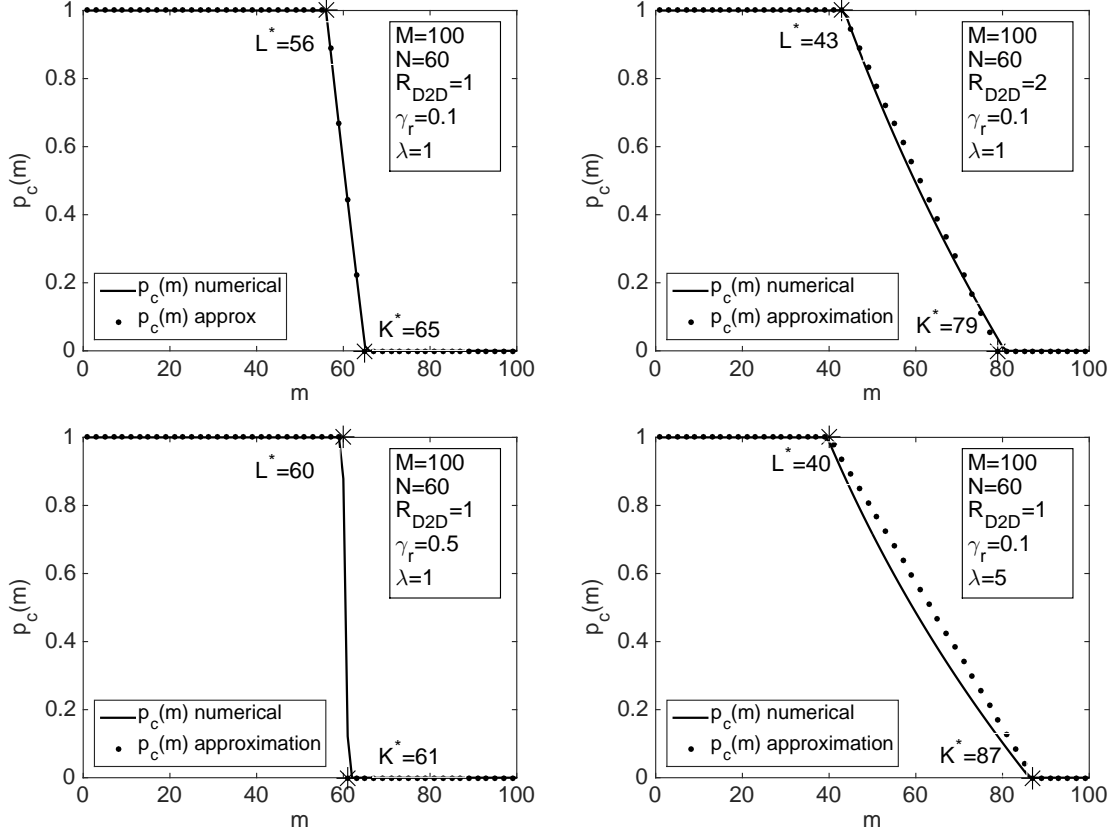


Fig. 7: Optimal cache placement (independently at each user) with more focused content popularity.

VIII. NUMERICAL COMPARISON OF DIFFERENT CONTENT PLACEMENT MODELS

In Sect. IV, we studied an exchangeable placement strategy and showed that it yields a positively correlated placement and is suboptimal. In Sect. V, we discussed GPPs modeling the pairwise interactions between users and showed that repulsion, negatively correlated placement techniques, e.g., F-D and MHC-inspired models, can provide a higher cache hit than independent placement.

In this section, we first present the spatial content placement results from Sects. III-VII. We compare the optimal solution $p_c^*(m)$ (5) and our linear approximation (9) in Fig. 7. Modifying the D2D parameters, we observe that our linear solution in (9) is indeed a good approximation of the optimal solution in (5). Keeping γ_r constant, by increasing R_{D2D} , we expect to see a more diverse set of requests from the user, L to decrease and K to increase. The converse is also true. When we keep R_{D2D} fixed, and increase γ_r , since the requests become more skewed

towards the most popular files, the optimal strategy for the user is to store the most popular files in its cache. Keeping R_{D2D} and γ_r fixed, and increasing λ has a similar effect as increasing R_{D2D} . From these plots, although it is clear that independent placement favors the most popular contents, it is not always optimal to cache the most popular contents everywhere.

The performance of the independent content placement and the MHC-based model is mainly determined by the cache size. Hence, the analysis boils down to finding the critical cache size that determines which model outperforms the other in terms of the hit probability under or above the critical size. Using the hit probabilities given in (3) and (37), respectively for the independent and MHC content placements, the required condition for which the MHC model performs better than independent placement is $\sum_{m=1}^M p_r(m) \mathbb{P}(\tilde{C}_m = 1) \geq 1 - \sum_{m=1}^M p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N} = k) (1 - p_c(m))^k$. A sufficient condition for this to be valid is given as

$$\mathbb{P}(\tilde{C}_m = 1) \geq 1 - \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N} = k) (1 - p_c(m))^k, \quad (46)$$

equivalent to the condition $e^{-\lambda_t \pi r_m^2} \leq e^{-p_c(m) \lambda_t \pi R_{D2D}^2}$.

Compared to the independent content placement model in [12], the MHC-A content placement model provides a higher cache hit probability for the small cache size regime, which we demonstrate next.

Now, we consider two regimes controlled by the cache size N . In the regime where MHC-A placement is better than independent placement, using (46), r_m is lower bounded as $\sqrt{p_c(m)} R_{D2D} \leq r_m$, for all m , and the volume fraction is lower bounded by $1 - \exp(-\lambda_t \pi p_c(m) R_{D2D}^2)$. Since a high exclusion radius is required for small cache size, MHC-A placement performs better than the independent placement for small cache size. When $r_m < \sqrt{p_c(m)} R_{D2D}$, the volume fraction is upper bounded by $1 - \exp(-\lambda_t \pi p_c(m) R_{D2D}^2)$. In this case, the file exclusion radii are very small for files with very low popularity, implying that the cache size should be sufficiently large, for which case independent placement is better than MHC-A placement. The cache hit probability trends of the independent placement in [12], and the MHC-A placement model with respect to the cache size are shown in Fig. 8. Note that the MHC-A model performs well given the cache size is small. We next show that the MHC-B model compensates the MHC-A model at the cost of communication radius.

From Fig. 9, we observe that the average hit probability for both cases improves with cache size

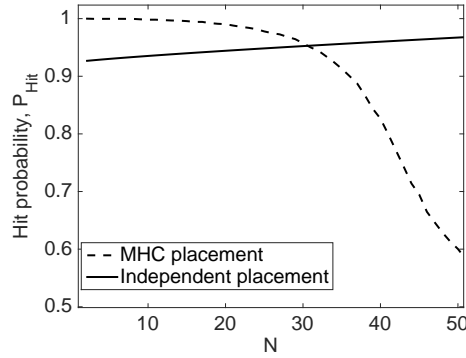


Fig. 8: Cache hit probabilities of the independent and MHC-A models.

N , independent placement improves with increasing R_{D2D} , and MHC-B is better than independent placement at low R_{D2D} . For small R_{D2D} , feasible for the D2D regime, MHC-inspired approaches are a better alternative. However, one disadvantage of the MHC models is that the excluded files' cache space is not reused, which can be resolved by jointly assigning marks. Therefore, we need to vectorize the marks to jointly determine the set of cached files and to avoid the problems caused by cache overuse. The calculation of the cache overuse probability is left as future work.

IX. CONCLUSIONS

We proposed spatially correlated content distribution models to maximize the hit probability by incorporating placement strategies to enable spatial diversity, e.g., spatially exchangeable cache model and Gibbs point process-based soft-core placement models that capture the pairwise interactions.

Our findings on spatial content caching suggest that the following design insights should enable more efficient caching models for D2D-enabled wireless networks:

Repulsive cache placement. Negatively correlated content placement rather than independent placement is required to maximize the cache hit probability. Due to the isotropy of the PPP process, we contemplate a rotation invariant caching model. To satisfy negative spatial correlation, geographical separation of the content within the neighborhood of a typical receiver is required. Therefore, in caching protocol design, it is important to incorporate an exclusion region around each cache, such that nodes inside the exclusion region are not allowed to cache simultaneously. We show that high cache hit rates in a PPP network can be achieved through a modified MHC

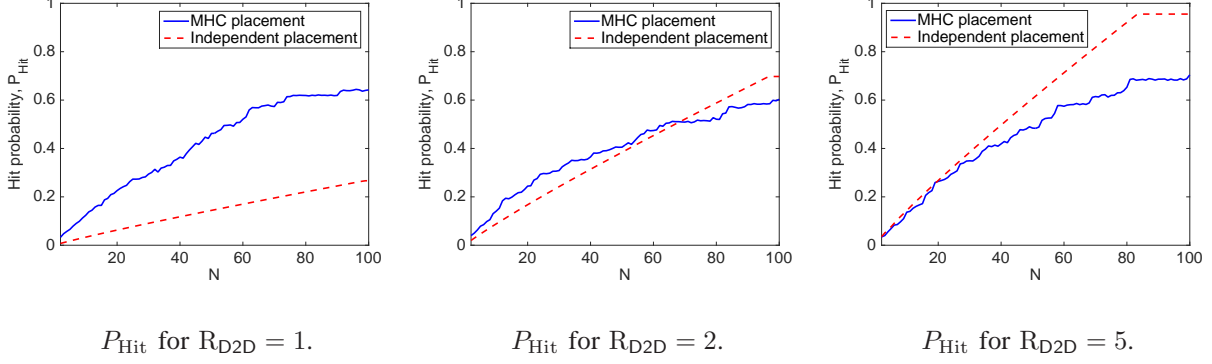


Fig. 9: MHC-B versus independent content placement.

placement model.

Hard-core and soft-core pairwise interactions, and Ising models. We analyzed a special case for GPPs, which is the MHC model, where the exclusions are determined by the hard-core radius, the general case with the soft-core placement incorporating pairwise correlations, i.e., Ising models, can yield more practical design insights, which we leave as future work. Using pairwise interactions, we can not only improve cache hit, but also the adaptiveness of our model. These interactions are promising because these potential functions characterize the temporal changes in the file popularities at different geographic locations, and can be exploited to develop caching policies that can easily adapt to the popularity dynamics. The shape and scale of the potential modeling the pairwise interactions should be determined accordingly. This can give insights into the spatial and temporal characteristics of a good caching policy and how fast it can adapt to geographical and temporal changes, and can pave the way for the development of spatial cache placement and eviction models to decide what content to discard, when to discard the content and where (to which neighbor) to relay the content.

Future studies include more general solutions for Ising or Gibbs models capturing the pairwise interactions, which can improve the cache hit via characterizing the temporal changes in the file popularities at different geographic locations, and can be exploited to develop adaptive caching policies. This can pave the way for the development of spatial cache placement and eviction models. Possible extensions also include hierarchical models for content delivery [20], multi-hop routing to improve the hit probability, distributed scheduling and content caching with bursty arrivals and delay constraints, and smoothing the cellular traffic by minimizing the peak-to-

average traffic ratio with D2D transmissions.

ACKNOWLEDGEMENT

Authors thank Abishek Sankararaman for many helpful discussions on GPPs and Ising models.

REFERENCES

- [1] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing the spatial content caching distribution for Device-to-Device communications," in *Proc., IEEE ISIT*, Barcelona, Spain, Jul. 2016.
- [2] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40–48, Apr. 2014.
- [3] N. Naderializadeh, D. T. Kao, and A. S. Avestimehr, "How to utilize caching to improve spectral efficiency in Device-to-Device wireless networks," in *Proc., Annu. Allerton Conf.*, Illinois, USA, Oct. 2014.
- [4] "Cisco visual networking index: Global mobile data traffic forecast update, 20152020 white paper," white paper, Feb. 2016.
- [5] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Proactive content download and user demand shaping for data networks," *IEEE/ACM Trans. Netw.*, vol. 23, no. 6, pp. 1917–1930, Dec. 2015.
- [6] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, pp. 2856–67, 2014.
- [7] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.
- [8] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for Device-to-Device communications with distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286–4298, Jul. 2014.
- [9] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," *to appear, IEEE Trans. Commun.*, 2016.
- [10] E. Altman, K. Avrachenkov, and J. Goseling, "Coding for caches in the plane," *arXiv preprint arXiv:1309.0604*, 2013.
- [11] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, Dec. 2013.
- [12] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc., IEEE ICC*, London, UK, 2015.
- [13] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Modeling and performance analysis of clustered device-to-device networks," *IEEE Trans. Wireless Commun.*, *to appear*, 2016.
- [14] M. Afshang and H. S. Dhillon, "Fundamentals of modeling finite wireless networks using binomial point process," *arXiv preprint arXiv:1606.04405*, 2016.
- [15] N. Golrezaei, K. Shanmugam, A. Dimakis, A. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc., IEEE Infocom*, Mar. 2012.
- [16] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82 – 89, Aug 2014.
- [17] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Proactive data download and demand shaping," Ohio State University, Tech. Rep., 2013. [Online]. Available: <http://www2.ece.ohio-state.edu/~tadrousj/ProactiveTechReport.pdf>

- [18] A. Sengupta, S. Amuru, R. Tandon, R. Buehrer, and T. Clancy, "Learning distributed caching strategies in small cell networks," in *Proc., 11th International Symposium on Wireless Communications Systems (ISWCS)*, Barcelona, Aug. 2014, pp. 917 – 921.
- [19] M. Leconte, G. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas, "Placing dynamic content in caches with small population," in *Proc., IEEE Infocom*, 2016.
- [20] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: Modeling, design and experimental results," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 7, pp. 1305–1314, Sep. 2002.
- [21] A. Giovanidis and A. Avranas, "Spatial multi-LRU caching for wireless networks with coverage overlaps," in *Proc., ACM SIGMETRICS/IFIP Performance*, Antibes, France, 2016.
- [22] H. P. Keeler, B. Błaszczyszyn, and M. Karray, "SINR-based k-coverage probability in cellular networks with arbitrary shadowing," in *Proc., IEEE ISIT*, Istanbul, July 2013, pp. 1167 – 1171.
- [23] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Trans. Wireless Comm.*, vol. 13, no. 12, pp. 6727–6740, Dec. 2014.
- [24] F. Baccelli and B. Błaszczyszyn, *Stochastic Geometry and Wireless Networks*. NOW: Found. Trends. Network., 2010.
- [25] D. Aldous. (2013) Exchangeability and related topics. [Online]. Available: www.stat.berkeley.edu/~aldous/206-Exch/
- [26] J. Møller and R. P. Waagepetersen, *Statistical inference and simulation for spatial point processes*. Boca Raton: Chapman and Hall/CRC, 2004.
- [27] D. Ruelle, *Statistical Mechanics*. New York: John Wiley & Sons, Inc., 1969.
- [28] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 2nd ed. John Wiley and Sons, 1996.
- [29] B. A. Cipra, "An introduction to the ising model," *American Mathematical Monthly*, vol. 94, no. 10, pp. 937–959, 1987.
- [30] G. Mie, "Zur kinetischen Theorie der einatomigen Körper," *Annalen der Physik*, vol. 316, no. 8, pp. 657–697, 1903.
- [31] J. E. Jones, "On the determination of molecular fields. II. From the equation of state of a gas," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 106, no. 738, 1924, the Royal Society.
- [32] R. J. L. Roy, N. S. Dattani, J. A. Coxon, A. J. Ross, P. Crozet, and C. Linton, "Accurate analytic potentials for Li₂ (X Σ 1g⁺) and Li₂ (A Σ 1u⁺) from 2 to 90 Å, and the radiative lifetime of Li (2p)," *The Journal of chemical physics*, vol. 131, no. 20, pp. 204–309, 2009, aIP Publishing.
- [33] Y. Ogata and M. Tanemura, "Estimation of interaction potentials of spatial point-patterns through the maximum-likelihood procedure," *Ann. Inst. Statist. Math.*, vol. 33, p. 31538, 1981.
- [34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2009.
- [35] M. Gerasimov, V. Kruglov, and A. Volodin, "On negatively associated random variables," *Lobachevskii Journal of Mathematics*, vol. 33, no. 1, pp. 47–5, 2012.
- [36] D. P. Dubhashi, V. Priebe, and D. Ranjan, "Negative dependence through the FKG inequality," *BRICS Report Series*, vol. 3, no. 27, 1996.
- [37] D. P. Dubhashi and D. Ranjan, "Balls and bins: A study in negative dependence," *BRICS Report Series*, vol. 3, 1996.
- [38] F. Reif, *Statistical thermal physics*. Mcgraw-Hill Kogakusha, 1965.
- [39] J. S. Blakemore, *Semiconductor statistics*. Courier Corporation, 2002.
- [40] M. Massimi, *Pauli's exclusion principle: The origin and validation of a scientific principle*. Cambridge University Press, 2005.
- [41] R. B. Leighton, *Principles of modern physics*. New York: McGraw-Hill, 1959, vol. 795.

- [42] M. Haenggi, *Stochastic geometry for wireless networks*. Cambridge University Press, 2012.
- [43] —, “Mean interference in hard-core wireless networks,” *IEEE Commun. Lett.*, vol. 15, no. 8, pp. 792–794, 2011.
- [44] J. Møller, M. L. Huber, and R. L. Wolpert, “Perfect simulation and moment properties for the Matérn type III process,” *Stoch. Process. Appl.*, vol. 120, pp. 2142–58, 2010.
- [45] M. Hörig and C. Redenbach, “The maximum volume hard subset model for Poisson processes: simulation aspects,” *J. Statist. Comput. Simul.*, vol. 82, p. 10721, 2012.